Fingerprinting Past the Front Page: Identifying Keywords in Search Queries over Tor

Se Eun Oh   Nicholas Hopper
University of Minnesota

Abstract

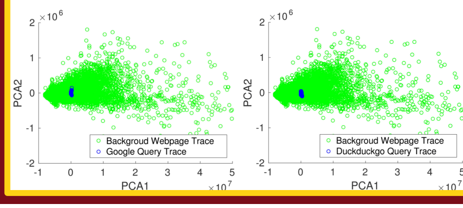
- In this work, we introduce a Keyword Fingerprinting (KF), extending Website Fingerprinting (WF), to identify keywords in search queries. Based on a two-stage, traffic analysis-based approach with new task-specific feature sets, a passive network adversary can defeat the use of Tor.
- We demonstrate the feasibility of the KF attacks across four popular search engines and various experimental settings (e.g., user query setting). We also further explore why several keywords are better fingerprintable.

Keyword Fingerprinting (KF)

- The attacker will progress through two sequential fingerprinting steps.
  - 1st step: Webpage fingerprinting to identify the query result traffic of the specific search engine
  - 2nd step: KF to predict keywords in query traces by both binary and multi-class classification
- KF focuses on 2nd step, which is challenging for existing WF techniques.

KF vs. WF

- CUMUL classifiers proposed by Panchenko et al. perform very well for the 1st step, which detects blue against green area. However, when identifying and differentiating keywords in blue, classifiers based on WF features perform poorly.



RESP feature set

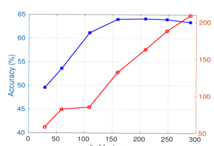
- All 80,000 query traces included a long sequence of incoming packets at the end of the trace. We call it "Resp" and remaining portion "Request".
- **Resp** is more informative than the request portion

Metric	Google		DuckDuckgo	
	RQ	RP	RQ	RP
Avg of # of packets	140	223	102	193
Max # of packets	288	559	251	801
Avg of total payload(KB)	115	496	89	434
Max of total payload(KB)	350	1246	295	1669
SVM Accuracy(%)	13.88	17.22	14.69	20.83

- We extracted **Resp** feature sets; Total number of TLS records, max, mean, sum of TLS record sizes (**RespTotal**); Sequence of cumulated size of TLS records (**cumulRespTLS**); Sequence of the corresponding number of Tor cells (**cumulRespTorCell**)

Data Preparation

- Reverse cumulRespTLS and cumulRespTorCell
  - The last elements are total size of TLS records and total number of Tor cells in Resp and good features to identify search terms
  - SVM accuracy for the first and last 140 packets in cumulRespTLS: 21.33% vs. 53.79%
- Number of Features: Use 247 features as it gave the best accuracy as well as acceptable running time



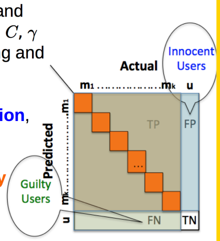
Feature evaluation using λ^2 statistics

- We tested different combinations of feature sets whose λ^2 statistics was higher than 6,000 and the best feature set was "**Aggr4**" aggregating Total, RespTotal, RcumulRespTLS, and RcumulRespTorCell

Feature	SS	MS	χ^2
roundedTCP	4.5e+10	4.55e+8	1353
roundedTLS	6.35e+10	6.42e+8	1905
cumulTLS	7.08e+10	7.15e+8	2123
Total	2.15e+11	2.17e+9	6461
burstIncoming	2.8e+11	2.83e+9	8402
RcumulRespTLS	2.22e+11	2.24e+9	6667
RcumulRespTorCell	2.17e+11	2.19e+9	6528

Support Vector Machine

- We used a non-linear classifier with a radial basis function (RBF) and 10-fold cross validation to find C, γ and to split dataset into training and testing set.
- Metrics
  - Binary Classification: **Precision**, **Recall** (TPR), FPR (%)
  - Multi-class classification: **Within-monitored Accuracy** (WM-acc) (%)



Confusion matrix for multiple outcomes

TPR and FPR when we identify 10k Google and Duckduckgo query traces against 100k webpage traces

- Google query trace identification

Ratio	0.1	0.2	0.3	0.5	0.8
TPR(%)	99.82	99.82	99.95	99.84	99.84
FPR(%)	0	0	0.0001	0.0001	0
precision(%)	100	100	99.98	99.99	100

- Duckduckgo query trace identification

Ratio	0.1	0.2	0.3	0.5	0.8
TPR(%)	99.94	99.94	99.96	99.94	99.94
FPR(%)	0	0	0	0	0
precision(%)	100	100	100	100	100

**Ratio means Monitored set size : Total set size

Closed and Open World Experiment

- Closed-world accuracy (10k keywords and 100 classes)

feature	Accuracy(%)
Total	35.48
torCell	7.54
roundedTCP	12.73
roundedTLS	15.16
burstIncoming	26.7
cumulTLS	18.67
RespTotal	26.14
RespTLS	17.22
RcumulRespTorCell	53.43
RcumulRespTLS	53.79
Aggr2	62.23
Aggr3	63.43
Aggr4	64.03

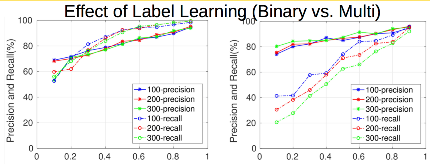
- Identifying 100 monitored keywords against 10k background keywords

Metric	Binary-label	Multi-label
TPR(%)	93.12	82.56
FPR(%)	14.88	8.09
Precision(%)	86.27	91.11

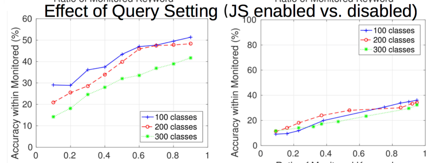
- Comparison to CUMUL classifier

Metric	cumulTLS	Aggr4
TPR(%)	34.95	82.56
FPR(%)	3.94	8.09
WM-Accuracy(%)	0.01	56.52

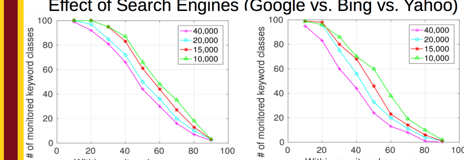
Effect of Label Learning (Binary vs. Multi)

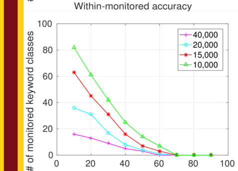


Effect of Query Setting (JS enabled vs. disabled)



Effect of Search Engines (Google vs. Bing vs. Yahoo)





Identifying 20k related searches using Binary classification

Metric	80ins
TPR	68.8
FPR	0.0005
Precision	99.85

TPR and Analysis on search result HTML

TPR(%)		# link	# domain	# Tag	# attribute
Google	0	49	10	845	1,575
	40	72	11	1,014	1,989
	80	84	14	1,378	2,749
Bing	0	33	10	406	533
	40	42	12	461	654
	80	118	18	826	1,410
Yahoo	0	46	1	527	928
	40	106	1	820	1211
	80	N/A	N/A	N/A	N/A

TPR(%)		max depth	# block	# tag direction change	len(HTML)	len(Data)
Google	0	24	37	244	128k	1,684
	40	30	49	319	165k	2,030
	80	35	62	449	232k	2,745
Bing	0	13	32	142	44k	807
	40	12	41	170	47k	914
	80	14	77	378	58k	1,635
Yahoo	0	18	30	191	92k	1,048
	40	20	63	390	96k	1,638
	80	N/A	N/A	N/A	N/A	N/A

** block=count # Blocks based on depth=18 for Google, 9 for Bing, and 14 for Yahoo, len()=number of characters